UNLIMITED

2

RSRE MEMORANDUM No. 4330

ROYAL SIGNALS & RADAR ESTABLISHMENT

EXPERIMENTS IN VARIABLE FRAME RATE ANALYSIS FOR SPEECH RECOGNITION

Authors: K M Ponting & S M Peeling

PROCUREMENT EXECUTIVE,
MINISTRY OF DEFENCE,
R S R E MALVERN,
WORCS.

RSRE MEMORANDUM No. 4330

UNLIMITED

0064316

CONDITIONS OF RELEASE

BR-113031

************************** DRIC U

************************** DRIC Y

Reports quoted are not necessarily available to members of the public or to commercial organisations.

A

Royal Signals and Radar Establishment

Memorandum 4330

Experiments in Variable Frame Rate Analysis for Speech Recognition

K.M. Ponting and S.M. Peeling

December 15, 1989

Abstract

The application of a simple variable frame rate analysis to the RSRE Airborne Reconnaissance Mission system, a continuous speech recognition system based on phone-level hidden Markov models, is described. Results are presented which show that, using standard three state models, the addition of the variable frame rate analysis results in considerably improved performance, which is close to that obtained using simple duration sensitive models.

Keywords; Great Britain. ([illegible])

THIS PAGE IS LEFT BLANK INTENTIONALLY

Contents

List of Figures

List of Tables

1 Introduction

Variable frame rate (VFR) techniques are an established method for data reduction in speech coding, but have not so far attracted much attention for improving automatic speech recognition.

In [1] variable frame rate analysis was shown to give improved performance over fixed frame rate analysis at the same (average) data rate, and equivalent performance to the use of a fixed frame rate analysis at twice that average rate. This reduction in the number of frames processed without loss of performance provides a considerable saving in computational load, and VFR analysis has been used in various speech recognition systems for that reason, (eg [3]).

This memo describes the application of a simple VFR analysis to the RSRE Airborne Reconnaissance Mission (ARM) system. This is a continuous speech recognition system based on phone-level hidden Markov models (HMMs) which has been developed at the RSRE Speech Research Unit.

HMMs, although very powerful for building statistical models of speech, are poor at modelling durations. A number of methods have been proposed to overcome this deficiency, none entirely satisfactory. However, VFR analysis not only reduces the average data rate but can also provide duration information in a form compatible with standard HMM algorithms.

This paper describes the effect of varying the VFR threshold on recognition performance, both for a standard HMM topology and for a simple duration sensitive topology (based on [10]).

2 The Variable Frame Rate Algorithm

This section will briefly describe the nature of the data, what VFR analysis is, and its application to automatic speech recognition.

Assume that at any "instant" in time the speech signal can be represented by an ordered set of numbers, or feature vector. This "instant" is assumed to be short enough that the properties of the speech signal do not change significantly. Any utterance, or collection of words, can then be described as a succession of feature vectors (sometimes referred to as frames). There are areas where the speech signal is relatively constant and hence successive feature vectors will be very similar. In other areas the signal may change rapidly and hence successive feature vectors will be different.

In order to reduce the processing time, one obvious solution is to reduce the data (frame) rate. However, parts of the signal which are changing rapidly contain valuable information and so need to be retained. For this reason it is necessary to employ some method of data reduction which actually depends on the data. Variable frame rate coding is such a technique.

One of the first uses of VFR for data reduction in automatic speech recognition is described in [1]. In that paper the authors describe several different VFR algorithms. This memo has utilised the simplest of these.

The VFR algorithm has been designed to retain all the input feature vectors when they are changing most rapidly and omit a high proportion when they are relatively constant. A subset of the feature vectors is selected, thus avoiding the need for deciding how to combine vectors. All that is required is the calculation of some measure of similarity between two feature vectors and the comparison of this similarity measure with a threshold. The most common similarity measure used is the Euclidean distance which is used in all the experiments quoted here.

In the simple version of the algorithm the distance is computed between the last retained feature vector and the vector under consideration. The current vector is then omitted if the distance is less than the threshold. With this approach, a threshold less than the minimum distance (zero) results in vectors being retained at the original frame rate. A threshold set to the maximum distance (effectively infinity) would result in a single vector being output, and an intermediate threshold provides a variable frame rate dependent on the speech data.

Specifically, if $D(i,j)$ is the distance between the previously selected frame j and the current frame i, and the threshold is T, then the rule is to select frame i as the next output frame if :-

$$D(i,j) \geq T$$

In some applications different thresholds are used which decrease with time so the likelihood of outputting a frame increases with time. This application has used a single threshold but has set an upper bound of 50 (referred to as the duplication factor) on the number of frames which can be represented by any one output frame, thus effectively incorporating a time constraint. This limit is only reached in long periods of silence and is necessary to ensure that they are not completely reduced.

3 HMMs and the TI Topology

The ARM system will not be described in full here. Further details can be found in [12], [16].

The theory and use of sub-word hidden Markov models for automatic speech recognition is now well established (eg [7]). These systems typically have distinct models corresponding to each phoneme in the language, which are combined according to a pronunciation dictionary to give whole word models for recognition. A large set of models is usually used, allowing different models for a given phoneme according to its immediate phoneme context (so-called triphone models).

The version of the ARM system used for the experiments in this paper used a smaller set of models: four models for non-speech sounds; six models of short common words[1] and sixty-one models of the phonemes in the ARM dictionary (some phonemes have two distinct models, for syllable-initial and syllable-final consonants, which is the only context sensitivity embodied in this model set).

All state output probability density functions of HMMs in the system are Gaussian with a diagonal (co)variance matrix.

[1] of, or, in, at, air, oh (used instead of sero sometimes)

Initial estimates of HMM parameters were obtained from a small quantity of hand labelled speech. Standard HMM algorithms were then used to train the models on a set of 36 ARM reports (224 sentences, 1985 words), using only sentence level labelling.

Two model topologies were used, with and without VFR analysis. Above a certain threshold the standard (three states per phoneme) topology became difficult to initialise, so the full range of VFR thresholds was investigated using a simple duration sensitive topology (the "TI" topology). This also allows some assessment of the implied duration model, provided by the inclusion of the VFR count as an extra parameter (in the feature vectors), to be modelled.

3.1 "Standard" Topology

For the standard HMM topology, the number of states in each model was as follows:

- one state for each non-speech model
- three states for each phoneme model
- three or six states for each short word model (according to whether the baseform phonemic transcription had one or two phonemes)

In all cases the only transitions allowed from a given state are loop (repeat the state) and forward (step to the next state). All paths through this simple progressive model must pass through all states.

3.2 TI Topology

Direct modelling of durations can be incorporated into the model building and speech recognition algorithms, but at considerable cost in training and recognition time (eg [14]). Various indirect methods have also been proposed (eg [13],[15]). However the simplest method of allowing for differing durations is to set the number of states in the HMM according to the average duration of the word or phoneme being modelled. This has been shown to provide improved performance over a fixed number of states, but at the expense of increasing the number of parameters to be estimated and the workload during both training and recognition ([10]).

This duration sensitive topology is referred to as the TI topology in this paper. For this topology, the number of states is estimated as follows:

1. train a set of standard models
2. use those standard models to provide an automatic segmentation of the training data, based on the known sentence transcriptions
3. calculate average phoneme and short word durations from the automatic segmentation
4. set the number of states for the phoneme and short word models equal to the average duration for that phoneme or word

Without VFR, this results in models with more than ten states for most phonemes in this application.

In order to allow for very short occurrences of the phonemes, not only loop and forward transitions were allowed, but also skip, omitting a single state (the so-called Bakis topology). Model entry transitions were allowed to the first two states and exit transitions from the last two states.

These models were initialised on a small set of hand labelled data. If any of the hand labelled occurrences of a phoneme were too short (i.e. less than half the average duration), the set of allowed initial and final states for that model was increased as required.

Finally, the loop and skip transition probabilities for each state of each phoneme were calculated by finding the average frame-state distance along the correct paths in the training data, and setting the probabilities so that the "penalty" for a loop or a skip was equal to this average distance, ϵ. Performance does not seem to be critically dependent on this value. All results reported here use the common value $\epsilon = 0.023077$. [2]

The primary reason for adopting this topology is that it allows the automatic selection of numbers of states per phoneme (to match the average number of frames) without applying the artificial lower limit of three states (and hence three frames using the standard topology) per phoneme. This becomes particularly important for high VFR thresholds.

4 Speech Representations

The speech data used were obtained by passing digitised speech signals through a 27 channel filter bank analyser at 100 frames per second. The filters are spaced on a non-linear frequency scale based on that in [5].

As specified in [12] all the HMMs used data which had the speech spectra replaced by a cepstral representation. The data used in this report consisted of either 12 or 16 mel frequency cosine coefficients (MFCCs) with an (extra) overall amplitude term.

The results in [12] report that performance improvements are obtained by using 12 MFCCs with time differences. These differences are essentially a method of including extra information about the surrounding frames. They consist of the differences between corresponding MFCC coefficients (and amplitude) at t+20 and t-20 milliseconds.

For most of the results reported here, the count of frames represented by a given feature vector after VFR analysis is appended to the vector as an additional feature. This results in 18 features for 16 MFCCs and 27 features for 12 MFCCs with time differences. The use of this additional feature provides a crude duration model, in that the model mean count, for a given state, will reflect the average number of frames in the original analysis which are condensed to a single frame corresponding to that model state.

This is not a simple duration model, as the Gaussian distribution of count values assumed by the model building and recognition is convolved with the geometric distribution of state

[2] The resulting system is very similar to a template based recogniser, with a more sophisticated method of obtaining templates, and "time distortion penalties" based on ϵ.

dwell time common to hidden Markov models. (The Gaussian assumption for this feature is even less valid than it is for the MFCC features). [3]

When time differences are used in conjunction with VFR analysis, the VFR analysis is done first, then time differences are taken between preceding and succeeding output frames (excluding the count features) and appended to the current output frame. The frames thus differenced could be up to 1 second apart, although in speech regions the maximum time difference observed is around 200mS.

5 Recognition and Scoring

The recognition algorithm used is a sub-word model implementation of a one-pass dynamic programming algorithm ([2]).

Whole reports are processed including silences between sentences.

Scoring is based on a dynamic programming alignment at the phoneme level, taking account of the known sentence end times, with subsequent marking of words according to whether their constituent phonemes correctly line up (cf [6], [11]).

Recognition results are reported for two levels of syntactic constraint. All the phone results come from employing the *simple* syntax in which any sequence of phonemes can be recognised. The word results are obtained from the *word* syntax which allows recognition of any sequence of non-speech sounds and words from the ARM vocabulary.

As with the results quoted in [12], these results are presented in terms of *% words correct* and *% word accuracy.* These are computed as follows, using dynamic programming to align the true transcription of the test data with the output of the recogniser:

$$\%\,words\,correct = \frac{N-S-D}{N} \times 100, \quad \%\,word\,accuracy = \frac{N-S-D-I}{N} \times 100$$

where N is the number of words in the test set, and S, D and I are the number of words substituted (i.e. recognised as the incorrect word), deleted and inserted respectively. The more interesting results are those in the columns headed "word accuracy" since these reflect more closely the level of performance which would be perceived by a user of the system.

As in [12] the training and test data were distinct ARM reports. Unless otherwise stated, all the recognition results are for a 540 word test set.

6 Results for Sixteen MFCCs

6.1 Effect of Different Thresholds on Data Files

When applying the VFR technique to speech data it is useful to know what sort of data reduction is being obtained. In the ARM system training and testing files are dealt with

[3] It has been suggested ([8]) that this count may reflect rate of change rather than duration as long steady state regions will have larger counts, rapid transition regions will have small counts.

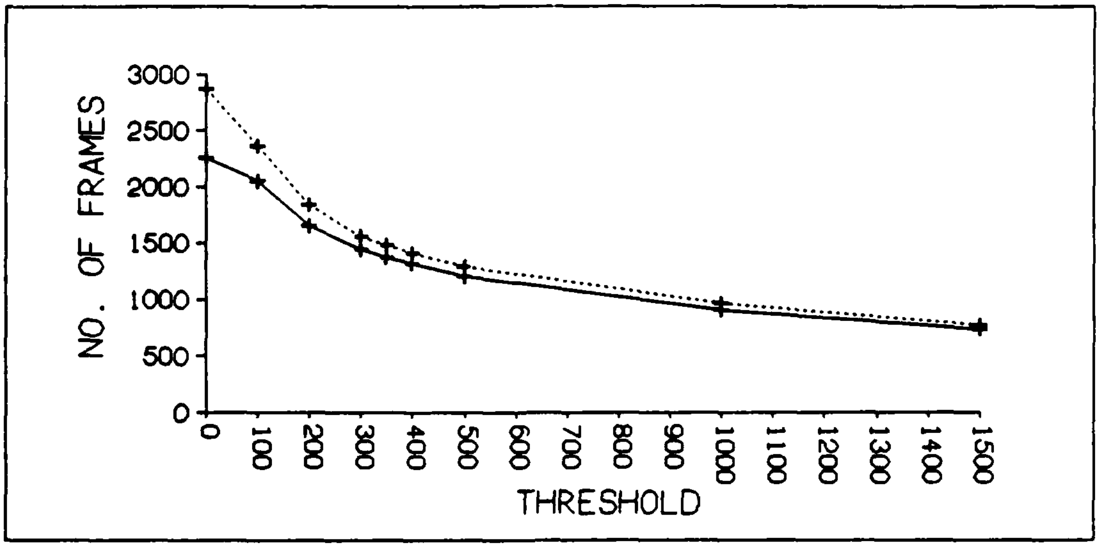

Figure 1: The effect of different thresholds on numbers of frames, from typical files, processed during training (solid line) and during testing (dotted line).

differently. For training purposes, only the actual speech in the file is used - any silence is ignored. During recognition however whole reports are processed, including silences (and breath noises etc) between sentences, unlike many other systems. Figure 1 shows the effect of the threshold on the number of frames processed for two typical files.

In this figure, the solid line shows the amount of speech used in a typical training file. The dotted line shows the total amount of speech (including silences) used in a typical testing file. The more rapid data reduction at low thresholds for the testing file is due to the silences being discarded. Notice that even a relatively small threshold (about 300-400) is capable of almost halving the amount of speech data to be processed. Also, after a threshold of 300, the rate of data reduction has begun to slow.

6.2 The First VFR Result

Threshold	Duration Information	Phone		Word	
		correct	accuracy	correct	accuracy
0	yes	63.1%	43.2%	83.5%	62.8%
350	yes	62.5%	49.9%	85.7%	70.7%
350	no	62.0%	51.2%	84.6%	68.7%

Table 1: Recognition results for standard three state models, 16 MFCCs, VFR thresholds as shown, with and without duration information.

In all the experiments described here, the ARM system used a limited set of hand-labelled phonemes to "seed" the inital models for the re-estimation process. The first experiments used three state models, so needed to ensure that after applying VFR analysis there were still at least three frames in all the hand labelled phones. The largest threshold which satisfied this criterion was 350.

Two experiments were then run. In the first a VFR threshold of zero was used (thus using data at the original frame rate) to give a benchmark against which future experiments could be compared. In the second experiment VFR analysis was applied with a threshold of 350. These recognition results are shown in Table 1 in the entries with duration information.

This result was unexpectedly good as the word accuracy using a threshold of 350 was significantly better than that obtained using the full data (ie the original frame rate). This was in contrast to the results reported in [1].

6.3 Duration Information

6.3.1 VFR Without Duration Information

By applying the VFR algorithm, data is both discarded and added: discarded in the sense that constant sections are removed and added in that the "count" now appears in each frame of data. In order to investigate which of these factors resulted in the improved recognition results shown in Table 1, a modified experiment was conducted. In this, the VFR analysis was performed in exactly the same way, but the count parameter was now no longer included in the data. In other words, the duration information has been discarded. These results are shown in Table 1 in the entry without duration information.

This result shows that most of the improvement comes from discarding frames, rather than the duration information. On the test set of 540 words a difference of 2% is not statistically significant. The experiment was therefore repeated using a much larger (and different) test set. The results for this appear in Table 2. The 2% improvement in word accuracy by including the count was maintained over this 2577 word test set. Comparing word accuracies as if they were proportions (a crude and inadequate measure, [4]), there is some evidence ($p = 0.6$) that this 2% improvement is more than just a random fluctuation.

Duration	Phone		Word	
Information	correct	accuracy	correct	accuracy
yes	63.8%	48.2%	85.8%	68.9%
no	62.7%	48.7%	84.7%	66.9%

Table 2: Recognition results over an increased test set for standard three state models, 16 MFCCs and a VFR threshold of 350, with and without duration information included.

On the basis of these results it was decided to continue to include the duration information in the data. All future results therefore apply to data which include the count parameter.

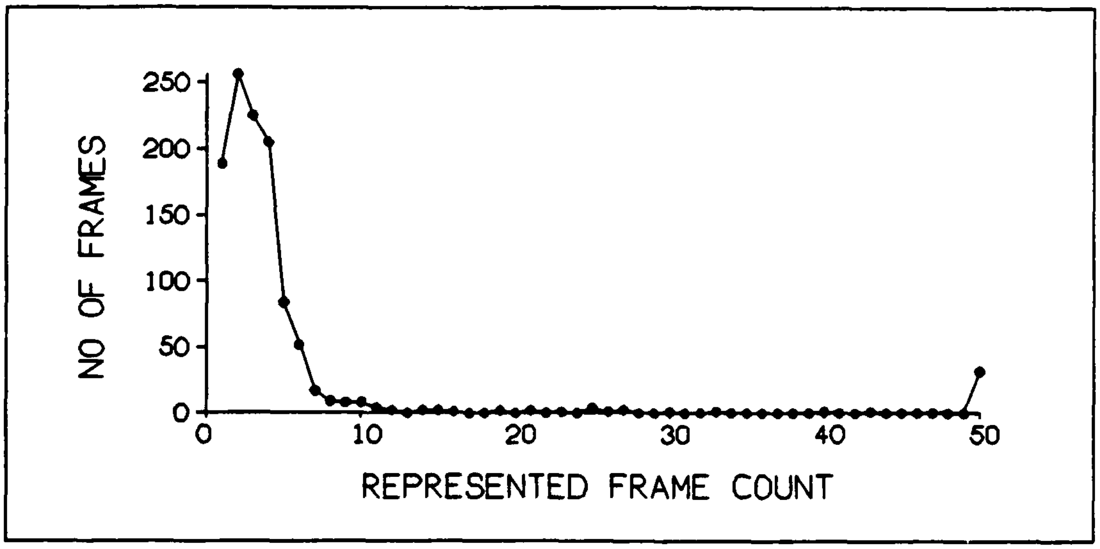

Figure 2: Distribution of frames represented by each VFR frame for a typical file with a VFR threshold of 1500.

6.3.2 Typical Duration Information

Figure 2 shows the behaviour of the count parameter over the 50 possible values for a data file which had used a VFR threshold of 1500. The "hump" at 50 is due entirely to periods of silence in the data file. As the threshold decreases this hump decreases in size. Similarly, since less data is discarded at lower thresholds, the peak near the axis is shifted closer to the axis.

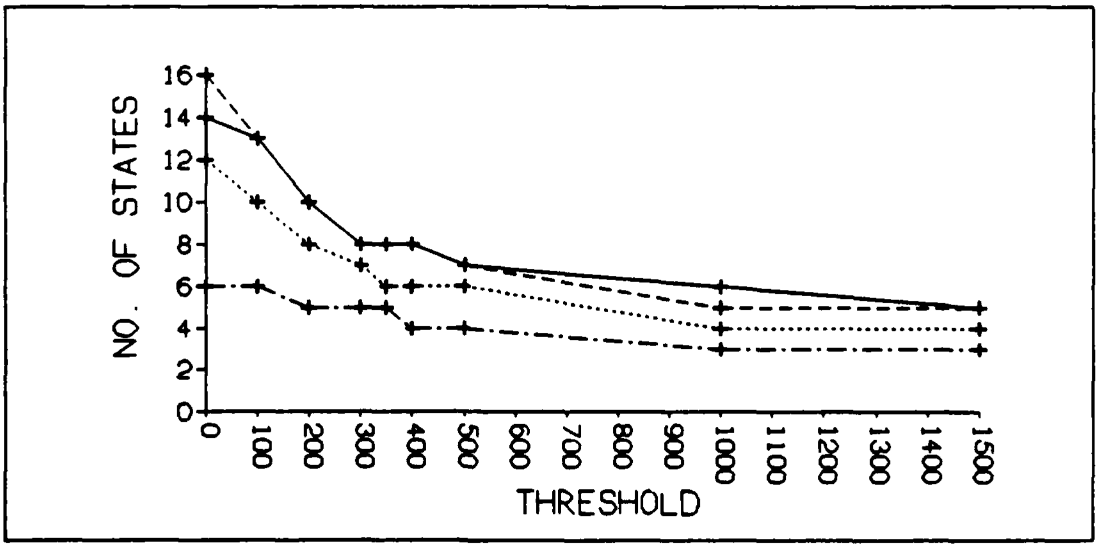

Figure 3: Variation of number of states with threshold for four phonemes: /aI/ - solid line; /O/ - dashed line; /i/ - dotted line and /k/ - dot-dash line.

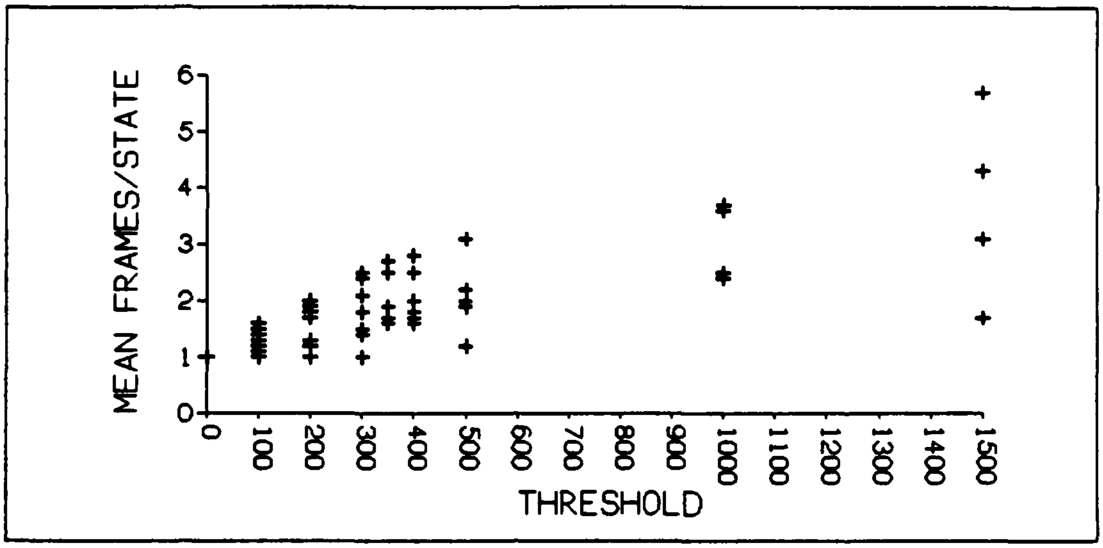

Figure 4: The distribution of the mean number of frames per state for the phoneme /i/.

6.4 Using the TI Topology

6.4.1 Effects of Using the TI Topology

In order to examine the recognition performance over a range of thresholds the "TI" topology, described in Section 3.2, was adopted. In this topology, the number of states for a phoneme is set to be the average duration of that phoneme in the training set.

Figure 3 shows how the number of states varies for different thresholds for four phonemes: /i/ as in heed, /aI/ as in hide, /O/ as in hoard and /k/ as in cake. An initial rapid decrease begins to level off at around threshold = 350. This behaviour has implications on processing time since training and recognition times are both approximately proportional to the number of states multiplied by number of speech frames.

Within models, as the threshold increases the number of states decreases and the number of original frames represented by each VFR frame increases. Hence, the number of original frames modelled by each state in the HMM will increase. This is reflected in the re-estimated values of the count element of each state mean, shown in Figure 4 for the phoneme /i/ and Figure 5 for /k/. [4]

For phoneme /k/ the distribution is fairly regular - the mean of the frames per state increases uniformly, whilst for /i/ the behaviour is less regular.

Figure 6 shows the behaviour of the count element at different positions in the phoneme /i/. The beginning and end represent the first and last states respectively. The middle state is either the true middle for an odd number of states or the average of the two centre

[4] Note that for /k/ there are models for the phoneme occurring at both the start and end of a word, the numbers in the graph represent both.

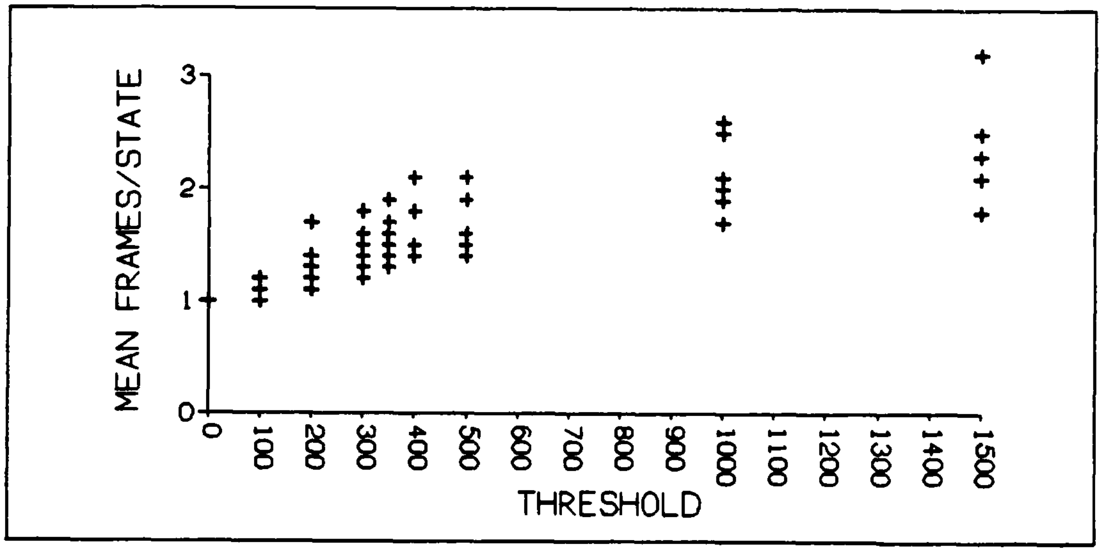

Figure 5: The distribution of the mean number of frames per state for the phoneme /k/.

values for an even number of states. As expected, this shows that the middle part of the phoneme contracts more with increasing threshold.

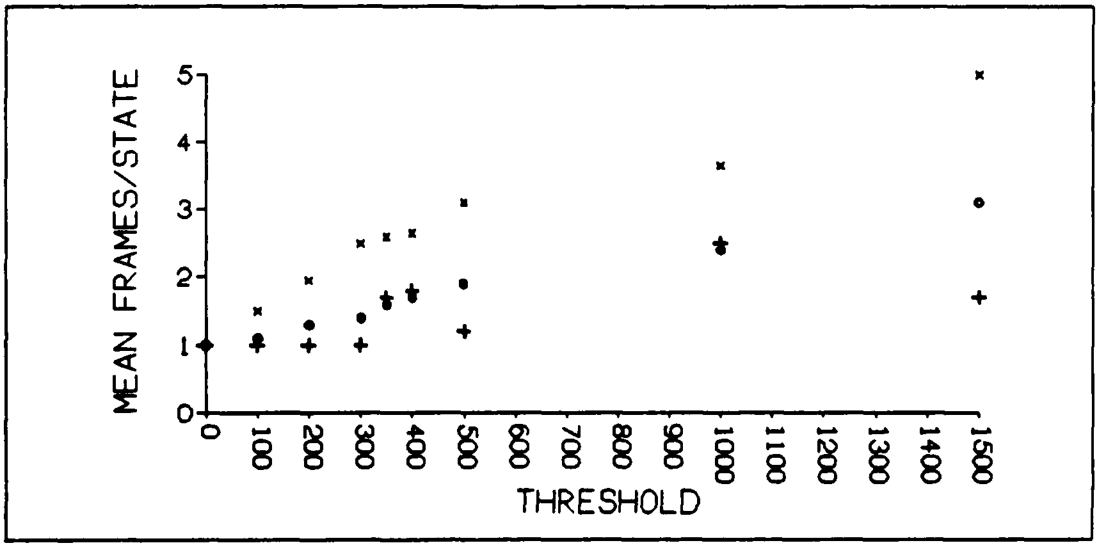

Figure 6: The distribution of the mean number of frames per state for the phoneme /i/ for different positions in the phone - plus for beginning; cross for middle and circle for end.

6.4.2 TI Topology Results

VFR	Phone		Word	
Threshold	correct	accuracy	correct	accuracy
0	66.0%	59.7%	85.7%	72.8%
100	66.1%	59.8%	86.1%	72.4%
200	65.4%	57.7%	85.7%	70.9%
300	65.7%	58.2%	85.6%	69.6%
350	65.2%	56.5%	86.7%	72.4%
400	65.3%	55.2%	86.3%	71.3%
500	64.5%	53.2%	85.9%	67.4%
1000	61.0%	44.2%	84.1%	58.3%
1500	57.6%	40.1%	80.0%	48.5%

Table 3: Recognition results for models using 16 MFCCs, the TI topology and VFR thresholds as shown.

The results of using the various VFR thresholds with 16 MFCCs are shown in Table 3 and Figure 7. On the full data set (i.e. a threshold of zero) there is a significant improvement in recognition performance by employing the TI topology. This is achieved at the expense of more states per model.

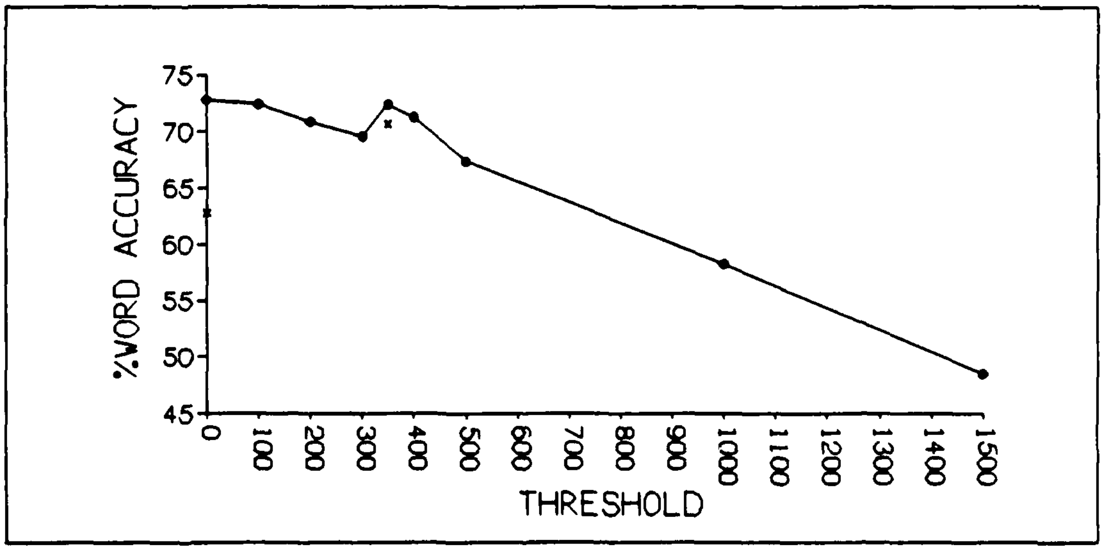

Figure 7: Percentage word accuracy over different VFR thresholds using 16 MFCCs and TI topology. Crosses represent results from using standard three state models.

In Table 3 the percentage of words correct are not significantly different over thresholds between 0 and 500. Hence, from the formulae in section 5, any decrease in accuracy over this range is due to a greater number of insertions.

From Figure 7 it can be seen that the results from using standard 3-state HMMs and a VFR threshold of 350 produces results nearly as good as the best obtained using the TI topology. This is very encouraging from an implementation point of view since there are many more states for each phoneme with the TI topology and hence greater computing time required for training and for recognition.

7 Results for Twelve MFCCs Plus Differences

Most of the above experiments were then repeated using 12 MFCCs plus differences.

7.1 Results Using VFR and Standard Models

VFR	Phone		Word	
Threshold	correct	accuracy	correct	accuracy
0	64.6%	40.0%	85.7%	58.0%
350	67.0%	58.2%	84.6%	69.1%

Table 4: Recognition results for standard three state models, 12 MFCCs plus differences and VFR thresholds as shown.

As before, two VFR thresholds were used with the standard three state models. Zero was again used to provide a benchmark, and 350 which had produced good results for the 16 MFCCs was also used. These results are shown in Table 4 and show similar behaviour to the corresponding ones for 16 MFCCs in Table 1. The only significant improvement, over 16 MFCCs, from using 12 MFCCs plus differences comes in the phone results with a threshold of 350. Although the two thresholds are not strictly equivalent, since the feature vectors are of different sizes, pilot experiments showed that strict equivalence was not crucial.

7.2 Results Using TI Topology

VFR	Phone		Word	
Threshold	correct	accuracy	correct	accuracy
0	69.8%	62.3%	88.0%	75.6%
100	69.8%	63.2%	87.4%	74.1%
200	70.0%	64.5%	86.9%	73.7%
300	69.8%	63.6%	85.7%	70.9%
350	70.0%	61.4%	86.5%	71.1%
400	69.6%	60.1%	86.3%	68.9%
500	68.5%	58.4%	87.0%	67.0%
1000	64.5%	49.6%	82.4%	52.4%
1500	62.6%	44.8%	79.1%	42.6%

Table 5: Recognition results for models using 12 MFCCs plus differences, the TI topology and VFR thresholds as shown.

The results using 12 MFCCs plus differences for various VFR thresholds, are shown in Table 5 and Figure 8.

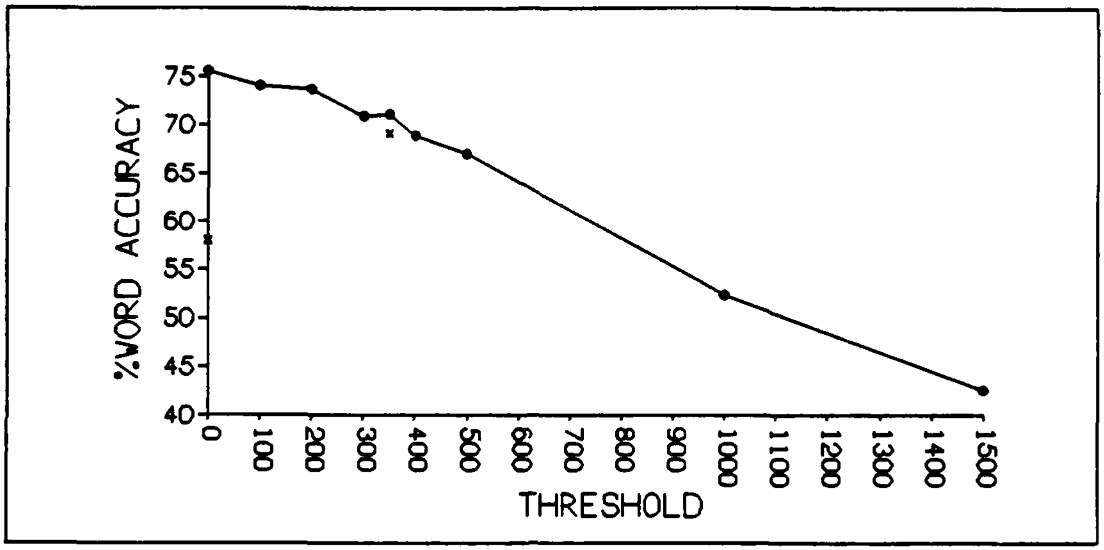

Figure 8: Percentage word accuracy over different VFR thresholds using 12 MFCCs plus differences and TI topology. Crosses represent results from using standard three state models.

Once again there is a significant improvement in performance at a threshold of 0 from using the extra states for each model with the TI topology.

Unlike the 16 MFCC results, 12 MFCCs plus differences give a steady decrease in performance and no improvement once VFR is employed.

7.3 Changing the Value of ϵ

Initial work [16] had suggested that the value of ϵ used by the TI topology was not crucial. Two experiments were conducted in which the value of ϵ was recalculated at the end of the first run and a new run executed using this reestimated value. These experiments used VFR thresholds of 1000 and 1500 with ϵ values of 0.063763 and 0.084961 respectively. Results are not reproduced here since they were virtually identical to those shown in Table 5.

8 Conclusions

It has been shown that VFR analysis can be used successfully for speech recognition. Within the ARM project VFR analysis has been shown to improve the recognition performance of the system. This is believed to be the first demonstration of this behaviour.

VFR analysis with a threshold of 350, and standard 3 state models, gives substantially improved word recognition results (over those using no VFR) when using either 16 MFCCs or 12 MFCCs plus differences.

Using the TI topology gives improved recognition results over those from standard three state models.

For 16 MFCCs using a VFR threshold of 350 and standard 3 state models produces results comparable to those obtained from using the TI topology.

Using the TI topology, the results from 16 MFCCs and 12 MFCCs plus differences are very similar. The phone accuracy from the latter tends to be better.

When using 16 MFCCs and the TI topology, there is a no significant degradation in performance from using a VFR threshold of up to 400. In the case of 12 MFCCs plus differences there is a steady decrease in word accuracy when employing VFR. In view of this and the increased processing time for 12 MFCCs plus differences, work in a companion paper ([9]) has concentrated on 16 MFCCs.

Results suggest that when using the TI topology, the choice of the value of the parameter ϵ is not crucial.

9 Future Work

It can be seen in Figure 1 that the decrease in the number of frames processed is markedly slowing around the threshold=350 mark. It is possible that this behaviour could be exploited in future work in order to discover a good starting value for a VFR threshold.

All these results were obtained using data initially analysed at 100 frames per second (which is standard throughout the ARM project). However, [1] suggested that the data should be analysed at 200 frames per second. Future work will investigate the effect of higher initial frame rates on the recognition performance obtained after VFR analysis.

Further experiments using VFR analysis within the ARM project are reported in [9]. This work concentrates on using 16 MFCCs with triphone models in which each phone is modelling according to the preceding and succeeding phoneme context.

References

[1] J S Bridle and M D Brown, "A Data-Adaptive Frame Rate Technique And Its Use In Automatic Speech Recognition", Proc. Institute of Acoustics Autumn Conf.,Bournemouth, UK, 9-10 November, 1982, pp C2.1-C2.6.

[2] J S Bridle, M D Brown and R M Chamberlain, "A One-Pass Algorithm for Connected Word Recognition", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Paris, 1982, pp899-902.

[3] Y L Chow, M O Dunham, O A Kimball, M A Krasner, G F Kubala, J Makhoul, P J Price, S Roucos and R M Schwartz, "BYBLOS: The BBN Countinuous Speech Recognition System", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, 1987, pp89–92.

[4] L Gillick and S J Cox, "Some Statistical Issues in the Comparison of Speech Recogntion Algorithms", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Glasgow, 23-26 May, 1989, pp532-535.

[5] J N Holmes, "The JSRU Channel Vocoder", IEE Proceedings, vol 127, Part F, number 1, February 1980, pp 53–60.

[6] M J Hunt, "Evaluating the Performance of Connected Word Speech Recognition Systems" Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, New York, 1988, pp457–460.

[7] K-F Lee, "Large Vocabulary Speaker-Independent Continuous Speech Recognition: the SPHINX System", PhD Thesis, Carnegie Mellon University, 1988.

[8] D B Paul, Private Communication.

[9] S M Peeling and K M Ponting, "Further Experiments in Variable Frame Rate Analysis for Speech Recognition", RSRE Memo 4336, 1990.

[10] J Picone, "On Modelling Duration in Context in Speech Recognition", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, New York, 1988, pp421-424.

[11] J Picone, G R Doddington and D S Pallett "Phone-Mediated Word Alignment for Speech Recognition Evaluation", unpublished draft ms., dated September 30, 1988.

[12] K M Ponting and M J Russell, "The ARM Project: Automatic Recognition of Spoken Airborne Reconnaissance Reports", to appear in Proceedings of 'Military and Government Speech Tech 89', Arlington VA, 13-15 November, 1989.

[13] L R Rabiner, B-H Juang, S E Levinson and M M Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities", AT&T Technical Journal, Vol 64, no 6, 1985, pp 1211–1234.

[14] M J Russell, "Maximum Likelihood Hidden Semi-Markov Model Estimation for Automatic Speech Recognition", RSRE Memo 3837, 1985.

[15] M J Russell and A E Cook, "Experimental Evaluation of Duration Modelling Techniques for Automatic Speech Recognition", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, 1987, pp2376–2379.

[16] M J Russell, K M Ponting, S M Peeling, S R Browning, R K Moore, I Galiano and P Howell, "The ARM Continuous Speech Recognition System" to appear in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Albuquerque, New Mexico, 1990.

THIS PAGE IS LEFT BLANK INTENTIONALLY

DOCUMENT CONTROL SHEET

Overall security classification of sheet ....................UNCLASSIFIED....................

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the box concerned must be marked to indicate the classification, eg (R), (C) or (S))

1. DRIC Reference (if known)	2. Originator's Reference Memo 4330	3. Agency Reference	4. Report Security Classification Unclassified

5. Originator's Code (if known) 7784000	6. Originator (Corporate Author) Name and Location ROYAL SIGNALS & RADAR ESTABLISHMENT ST ANDREWS ROAD, GREAT MALVERN WORCESTERSHIRE WR14 3PS
5a. Sponsoring Agency's Code (if known)	6a. Sponsoring Agency (Contract Authority) Name and Location

7. Title

EXPERIMENTS IN VARIABLE FRAME RATE ANALYSIS FOR SPEECH RECOGNITION

7a. Title in Foreign Language (in the case of Translations)

7b. Presented at (for Conference Papers): Title, Place and Date of Conference

8. Author 1: Surname, Initials PONTING K M	9a. Author 2 PEELING S M	9b. Authors 3, 4 . . .	10. Date 1989.12	pp. ref. 15

11. Contract Number	12. Period	13. Project	14. Other Reference

15. Distribution Statement

UNLIMITED

Descriptors (or Keywords)

Continue on separate piece of paper

Abstract

The applicationof a simple variable frame rate analysis to the RSRE Airborne Reconnaissance Mission system, a continuous speech recognition system based on phone-level hidden Markov models, is described. Results are presented which show that, using standard three state models, the addition of the variable frame rate analysis results in considerably improved performance, which is close to that obtained using simple duration sensitive models.

THIS PAGE IS LEFT BLANK INTENTIONALLY